Types of normative conflict and the effectiveness of punishment

Rauhut, Heiko ; Winter, Fabian

Abstract: While the current literature focuses on how social norms generate cooperation, the issue of norm-related conflict deserves more attention. We develop a new typology of normative conflict by combining Coleman's (1990) distinction between conjoint and disjoint norms with our own classification of commitment-related and content-related normative conflicts (Winter, Rauhut, and Helbing 2012). We outline a theory of how the four resulting types of normative conflict can be ordered. We provide real-life examples and typical game-theoretical conceptualizations of the four cases and suggest how they can be sorted according to their conflict potential and the extent to which conflict can be restored by punishment. We then discuss a prototypical laboratory study for each of the types, and show how our theoretical arguments can be applied. We conclude with a discussion of how previously anomalous empirical results can be re-thought and understood in light of our theoretical reasoning. Finally, we give suggestions for prospective empirical micro-level corroborations and for mechanism design.

DOI: https://doi.org/10.1515/9783110472974-012

Heiko Rauhut and Fabian Winter

Types of Normative Conflicts and the Effectiveness of Punishment

Abstract: While the current literature focuses on how social norms generate cooperation, the issue of norm-related conflict deserves more attention. We develop a new typology of normative conflict by combining Coleman's (1990) distinction between conjoint and disjoint norms with our own classification of commitment-related and content-related normative conflicts (Winter, Rauhut, and Helbing 2012). We outline a theory of how the four resulting types of normative conflict can be ordered. We provide real-life examples and typical game-theoretical conceptualizations of the four cases and suggest how they can be sorted according to their conflict potential and the extent to which conflict can be restored by punishment. We then discuss a prototypical laboratory study for each of the types, and show how our theoretical arguments can be applied. We conclude with a discussion of how previously anomalous empirical results can be re-thought and understood in light of our theoretical reasoning. Finally, we give suggestions for prospective empirical micro-level corroborations and for mechanism design.

1 Introduction

Social norms have a pivotal role in sociology. They can serve as a "lubricant" of social order and facilitate social interaction in coordination problems such as which side of the road to drive on, which greeting to use or what clothing to wear in which context. They can also solve cooperation problems by prescribing contributions to collective goods such as a clean environment, a safe neighborhood, or public infrastructure. Scholars of different schools of thought seem to converge around the idea that social norms emerge because they have positive consequences for society. In the functionalist approach, norms bridge the tension between individual self-interest and the functional prerequisites of society (Durkheim 1997; Parsons 1968; Dahrendorf 1977). The rational-choice literature also argues that norms emerge when there is a demand for them (Ullmann-Margalit 1977; Coleman 1990). A demand is typically given in situations where everybody has an interest that all others cooperate but oneself.

Note: We thank Nikos Nikiforakis and Hironori Otsubo for allowing us to reanalyze their data, the *Nature* publishing group for the right to reprint one figure from Fehr and Gächter (2002), and two anonymous reviewers for their helpful comments. Heikor Rauhut acknowledges support by the SNSF Starting Grant BSSGI0_155981. Correspondence should be addressed to HR or FW. Both authors contributed equally to this work and are listed intentionally in alphabetical order.

https://doi.org/10.1515/9783110472974-012

Mechanisms such as expected future interactions (Axelrod 1984), credible signals of long-term interests in mutual social exchange (Gambetta 2009), or reputation-seeking (Nowak and Sigmund 1998; Sigmund 2010; Wedekind and Milinski 2000; Berger and Rauhut 2014) can explain cooperative behavior even among rational egoists. Interestingly, the emphasis in the current literature is on the positive societal effects of social norms: "The view that norms are created to prevent negative externalities, or to promote positive ones, is virtually canonical in the rational choice literature" (Hechter and Opp 2001).

In contrast to the rich literature about the positive effects of social norms on cooperation, we concentrate on the largely neglected argument that social norms can also generate conflict. Members of the same group can hold profoundly different normative expectations of what ought to be done. This phenomenon, referred to as "normative conflict", generates conflict rather than cooperation. If ego holds a different norm to alter, she can do everything right and have the best intentions to cooperate, but nevertheless find that her behavior is conceived of as improper. They fall into conflict, despite both being convinced of having behaved adequately.

We start by introducing the concept of normative conflict by extending Coleman's (1990) conceptualization of norms. We give an introduction to social norms and cooperation, and exemplify how norms prescribe how target actors ought to behave to benefit the beneficiaries of the norm. We first focus on cases where all involved actors share the same norm. Normative conflict, in this case, is about the level of normative commitment: how much should each actor sacrifice her self-interest to comply with the norm? The second kind of normative conflict is about the normative content: which kind of behavior is prescribed or proscribed in a given situation? For example, people may hold exclusive norms of cooperation, such as equality versus equity norms.

Our main argument in this article is that punishment has different effects in these different types of normative conflicts. The standard case is the first type of conflict about the level of commitment. Here, research shows that punishment helps foster cooperation. Our idea is that if people agree which norm to follow, punishment typically helps push low contributors towards more cooperation. However, if people do not agree which behavior should be conducted, that is, which normative content should apply, punishment often has detrimental effects. In other words, if people disagree about the kind of normative behavior that should be followed, punishment leads to counter-punishment, feuds, and long-lasting conflicts. We develop our theoretical argument of the effectiveness of punishment based on real-life examples and evidence from experiments. We believe that our proposed typology of normative conflict is helpful in re-reading the evidence of norm enforcement, and that we shed new light on the question of when punishment is effective and when it is ineffective in promoting cooperation.

2 A typology of norm-related conflicts

Social norms define rules of how one ought to behave in a certain situation. To be more precise, in *norm-relevant situations*, almost every member of a population believes that almost every other member has certain behavioral expectations. This implies that norms are directed at certain actions, which can be called *focal actions* (Colman 1990:246).

The expectations about focal actions are directed towards *targets* of the norm (equivalently one may say *target actors* or *norm targets*). Target actors are defined by Coleman (1990:247) as follows: "For any norm, there is a certain class of actors whose actions or potential actions are the focal actions. [...] I will call members of such a class targets of the norm, or target actors." Most norms benefit a certain group of actors, who are called *beneficiaries* of the norm. These beneficiaries typically hold the norm and are potential sanctioners of the target actors. Coleman (1990:247) defines beneficiaries as "a class of actors who would benefit from the norm, potentially hold the norm, and are potential sanctioners of the target actors. These are actors who [...] assume the right to partially control the focal actions and are seen by others [...] to have this right." In summary, target actors are individuals who are forced to restrict their self-interest to follow the norm while beneficiaries are individuals who benefit from general adherence to this norm. Following the above definitions, we define *social norm* as follows. A *social norm* is a commonly shared behavioral expectation among beneficiaries and targets of a norm of how one ought to behave in a norm-relevant situation, which is enforced by sanctions in case of norm violations (see also Winter, Rauhut, and Helbing 2012).[1]

Whereas the current debate is dominated by the argument that social norms solve the problem of cooperation among rational egoists, we argue that social norms can also trigger conflict. In our view, conflict can emerge from two sources. First, conflict can emerge if target actors and beneficiaries of a norm belong to a different group with different interests. We call this *structural conflict*. Second, conflict can emerge if actors apply contradicting norms in the same norm-relevant situation. We call these conflicts *normative conflicts*, and distinguish commitment from content-related normative conflicts. In commitment-related conflicts, actors disagree about the extent to which the norms should restrain their self-interest. In content-related conflicts, actors disagree about which norm should be followed in which situation.

1 A related definition is suggested in the sociological tradition by Elster (1989:105): "A norm [...] is the propensity to feel shame and to anticipate sanctions by others at the thought of behaving in a certain, forbidden way. [...] This propensity becomes a social norm when and to the extent that it is shared with other people." In economics, similar definitions are used. Fehr and Gächter (2000:166) define a social norm as follows: "It is 1) a behavioral regularity; that is 2) based on a socially shared belief of how one ought to behave; which triggers 3) the enforcement of the prescribed behavior by informal social sanctions."

2.1 Structural conflicts

Much of the current literature focuses on the case of *conjoint norms*, where the beneficiaries and targets of the norm belong to the same set of actors.[2] In the case of doping, all athletes are the target and likewise benefit from the anti-doping norm. From an individual perspective, doping yields a relative advantage at the price of damaging one's health. Whereas many cyclists prefer to accept this price, the relative advantage vanishes if all athletes dope and end up with bad health, which is (paradoxically) the same relative position compared to the situation in which nobody dopes. As for many examples of conjoint norms, the social norm bridges the cleavage between self-interest and collective good and can be modeled as a prisoner's dilemma. The case of non-aggression norms in the trench warfare of the First World War represents another, by now classic, example of conjoint norms. Here, French and German soldiers reduced their mortality risk by complying with strong behavioral norms to conduct mutual fake assaults and show mutual respect of war interceptions (Ashworth 1980; Axelrod 1984).

For some norms, the targets of a norm and the beneficiaries fall apart. In this case of *disjoint norms*, the separation is typically associated with opposing interests and causes conflict instead of cooperation. We can observe such conflict of interest between parents as the beneficiaries of a certain norm and their children as the target of the norm. Coleman (1990:245) gives the example of a high school girl who is asked by her friends to join them in smoking marijuana. Whereas her friends disdain her reluctance, her parents disapprove her consent. In the area of gender differences, many norms are disjoint. Consider the norms that women should not pursue a profession, should not practice polygamous sex, or should not engage in politics. It seems that such norms are targeted towards women to the benefit of men. The conflict of interest between the beneficiaries and targets of a norm might even be more pronounced in the case of norms proscribing racial or homosexual discrimination.

We define *structural conflict* as the conflict of interest between the beneficiaries and targets of a disjoint norm. Both beneficiaries and targets share the same behavioral expectation of how one ought to behave in any given norm-relevant situation. Nevertheless, only the beneficiary profits from norm-compliant behavior, which is produced by the target of the norm at own cost (Figure 1). Structural conflicts do not necessarily depend on specific norms, but are an inherent property of some form of heterogeneity in a situation's social structure. Asymmetry between actors, like gender or a parent-child relationship, allow for diverging behavioral expectations and form a necessary condition for the emergence of structural conflict. Whether or not a structural conflict might exist can thus already be inferred by taking a close look at the actors and their current social context, even before considering their specific social norms.

2 The typology of conjoint and disjoint norms was introduced by Coleman (1990:247ff.).

Notes: The left image illustrates conjoint norms. All targets of the norm benefit from norm-compliant behavior. The right image illustrates disjoint norms that prescribe or proscribe certain behaviors of target actors, which benefit a different set of actors. The intermediate case between conjoint and disjoint norms is displayed in the middle.

Fig. 1: Structural conflicts by different types of norms (Source: Authors' compilation).

2.2 Normative conflicts

The specification of normative conflicts requires distinguishing two factors that generate behavioral expectations:[3] the kind of action that should be undertaken and the intensity of that action. We term the first element "normative content", defined as the kind of behavior that is prescribed or proscribed in a given situation. It provides information about which of the situation's characteristics should be evaluated when choosing an action. We term the second element "level of normative commitment". This indicates that social norms usually require an actor to restrict self-interest in favor of another person's or group's wellbeing. Consequently, we define this element as the extent to which an actor should sacrifice self-interest to comply with the norm. The level of normative commitment is not fixed. While some norms may require strong restrictions, others are less demanding.

The idea of content-related normative conflicts can be illustrated by the following examples. When it comes to performance-related salaries, blue-collar employees often consider harmful working conditions as an important determinant, while white-collar employees stress value creation (Hyman and Brough 1975). In another study, soldiers differed over whether military merits or the fact of being married with children ought to be considered important for deciding early demobilization after World War II (Stouffer 1949). Alternatively, a group of employees in a firm may call for equal pay in contrast to a second group demanding a payment scheme based on added value. Thus, attributes such as working conditions, family status or children may serve as normative "cues" which determine the allocation of scarce goods (such as money or demobilization).

Consequently, we define *normative conflict* as a transaction failure resulting from actors holding at least partially exclusive normative expectations. The distinction between content and commitment of a norm enables us to classify conflicts based on distinct contents versus distinct commitments. Normative conflicts are interesting inasmuch as they describe situations in which actors adhere to social norms, believe themselves to be behaving correctly, and nevertheless experience conflicts.

3 See also Winter, Rauhut, and Helbing 2012:920f.

Obviously, it is possible to imagine combinations of structural and normative conflicts. For example, there can be norm-relevant situations in which the same group of beneficiaries favors different disjoint norms, which would benefit them to the same extent. Thus, they do not agree on whether norm A or norm B is the appropriate norm that should be demanded from the norm targets to please the beneficiaries. Note that we concentrate only on the pure cases in this chapter.

3 The theory on the effectiveness of peer punishment for different types of norm-related conflicts

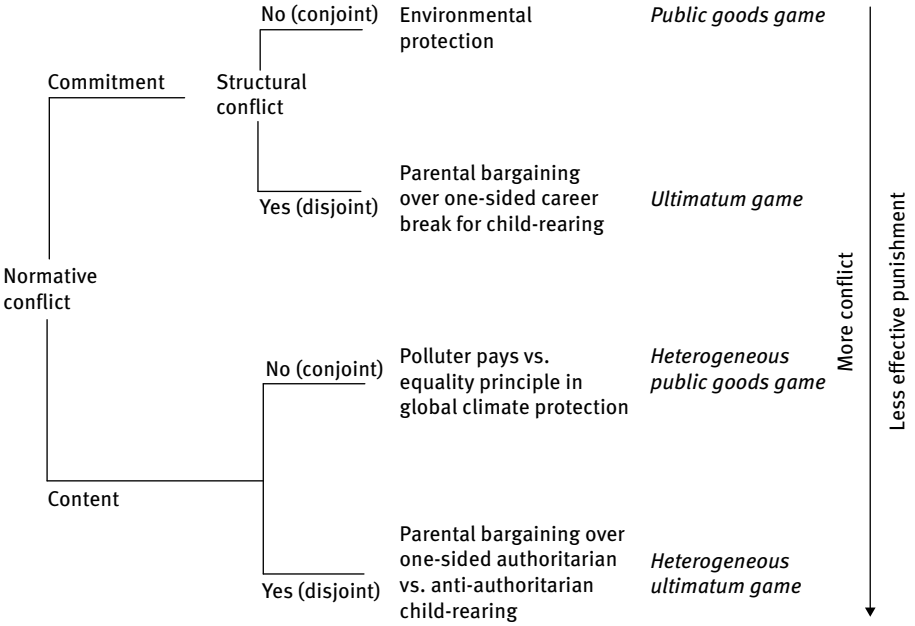
Our typology of structural and normative conflicts can be cross-tabulated two by two. This yields four cases. The cross-tabulation is depicted in Table 1 and visualized in Figure 2.

This typology is helpful in systematizing theoretical and empirical research on social norms. We illustrate this for schematizing research on the effectiveness of peer punishment for the promotion of cooperation norms. We conjecture that peer punishment is more effective for commitment-related than for content-related conflicts. It is also more effective in the absence of structural conflicts, where norms are conjoint rather than disjoint. Our reasoning suggests the following order for the effectiveness of punishment: commitment-related conflicts over conjoint norms, commitment-related conflicts over disjoint norms, content-related conflicts over conjoint norms, and then content-related conflicts over disjoint norms. This order is conceptualized by the arrow in Figure 2.

We illustrate our reasoning by giving examples and typical game-theoretical conceptualizations for each of the four cases. The first case of commitment-related conflicts over conjoint norms is the simplest and most prototypical one. A classic example is environmental protection (Diekmann and Preisendörfer 2003). All benefit if everybody contributes to environmental protection. The beneficiaries are also the target actors of the norm of eco-cooperation. The typical conflict is how much to contribute to eco-cooperation. In other words, people can disagree about the level of normative commitment. For example, is it sufficient to buy energy-saving lamps, or should one

Tab. 1: Effectiveness of punishment.

		Normative conflict	
		Commitment-related conflict	**Content-related conflict**
Structural conflict	No (conjoint norm)	Very high	Low
	Yes (disjoint norm)	High	Very low



Notes: This 2 × 2 typology yields four cases, for which typical examples are listed at each branch of the tree diagram. To the right of the examples are typical game-theoretical conceptualizations of the interaction structure. The four cases are ordered by increasing potential for normative conflicts and decreasing effectiveness of peer punishment. This order is conceptualized by the arrow on the right.

Fig. 2: Typology of normative conflicts by commitment versus content related conflicts when structural conflict is present or absent (Source: Authors' compilation).

also buy a fuel-thrifty car, or refrain from owning a personal car, or even abstain from flying to holiday destinations? A typical abstract conceptualization of these commitment-related conflicts is a public goods game where people can contribute more or less to a common pool, from which all group members benefit equally. Peer-punishment is most effective in this case, since it "only" coordinates the cooperation level.

The second case of commitment-related conflicts over disjoint norms can be illustrated by the example of parental bargaining over a one-sided career break for child-rearing. When expecting a child, a couple may be interested in one partner keeping to his or her career track to earn sufficient money for the family, while the other partner takes a career break to raise the child. In this case, one parent is the target of the norm and is expected to invest time and energy for child rearing, with the consequence of sacrificing some career advantages. The beneficiary, on the other hand, can continue his or her career. Normative expectations in this case are one-sided, so that this case satisfies the conditions of a disjoint norm. Beneficiary and target actor may bargain about how much the target actor should invest in child-rearing and how many career

options it is tolerable to lose. Disagreement may therefore emerge about the level of normative commitment the target actor is expected to fulfill.

The strategic interaction structure may be generalized to an abstract ultimatum game. A proposer can decide how to distribute a common pie and a responder can accept or reject. Rejection can be regarded as altruistic punishment, since the pie is lost to both parties. The structurally weaker responder often adheres to a fairness norm and rejects offers that are too low. This norm is disjoint, since target actor and beneficiary fall apart.

We expect punishment to be less effective for the enforcement of a requested level of commitment for the latter case of disjoint norms, compared to the former case of conjoint norms. The reason is that the conflict of interests in disjoint norms hampers the alignment of a mutually agreed level of commitment. In conjoint norms, there is no conflict of interests; both parties "merely" have to coordinate on how much self-interest should be restrained to benefit everybody in the group.

The third case of content-related conflicts over conjoint norms can be exemplified by distinct norms of environmental protection. Take the case of global climate protection. Some parties may argue that heavy polluters should contribute larger shares to global climate protection than low polluters. In contrast, other parties may adhere to an equality principle and may demand that all parties should contribute equally to global climate protection. This case exemplifies conjoint norms, since all target actors benefit equally from a cleaner and more protected global environment. However, target actors disagree about which normative contents should be followed to protect the environment.

In more abstract terms, the strategic interaction structure can be conceptualized by a heterogeneous public goods game. For example, target actors can have different production costs to produce the same level of the public good. To stay with our example, in countries with a lower technological level, the fulfillment of certain environmental guidelines takes higher relative prices compared to countries with a high technological level.[4] We expect punishment to be less effective here than in the former cases, since disagreement about normative principles is harder to resolve compared to disagreement about the level at which commonly agreed principles should be adhered to.

The fourth case of content-related conflicts over disjoint norms is illustrated by parental bargaining over different educational principles in a family that divides labor between child-rearing and breadwinning. This situation describes a disjoint norm, where the child-raiser is target actor of the norm to invest time and energy for child-rearing. The educator and the breadwinner may, however, disagree with the educa-

4 A comparable conflict over contents can be modeled by different *per capita* returns that target actors receive from the same production levels at same production costs from all contributing target actors (e.g., Nikifourakis). We will discuss this case in the next section.

tional principles. For example, one may favor an anti-authoritarian, and the other an authoritarian, style. The underlying motive of both styles may be similar inasmuch as both are geared towards making the best of the education of the child – they are just different means to serve this end.

In more abstract terms, the fourth case can be conceptualized by a heterogeneous ultimatum game. For example, a proposer and a responder can be heterogeneous in their contributions to a common pool, which needs to be divided. A high-contributing responder may demand more than equal shares from the common pool, while a low-contributing proposer has a structural advantage and may insist on equal shares. This yields a disjoint normative situation between proposer (beneficiary) and responder (target actor), where both adhere to different normative contents (an equity versus an equality norm). This represents an abstract model of a structurally advantaged breadwinner, who requests that the child-rearer follow his or her favored norm. We argue that the conflict is largest in this fourth case: there is disagreement about the content of the norm, and there is a structural conflict of interests between target actor and beneficiary. Therefore, we expect punishment to be least effective in these situations.

Comparing cases one and two with cases three and four, we expect disagreement about the level of commitment to be more easily resolvable than disagreement about normative contents. In commitment-related conflicts, everybody agrees about the normative principles. Punishment "merely" helps to align contribution levels in the group. In content-related conflicts, however, people disagree about which principle should be followed to produce the public good. This is a more fundamental conflict, where punishment is likely to provoke counter-punishment, feuds, and barely-resolvable cleavages. This reasoning leads us to the proposed order of the level of conflict and the effectiveness of punishment for the four types of norm-related conflicts.

One theoretical reason for the order of the level of conflict and the effectiveness of punishment is that the types can also be ordered by the number of potential conflicts. The number of potential conflicts is increasing from the first to the last type. Commitment-related conflicts over conjoint norms have one source of conflict: the level of commitment. Commitment-related conflicts over disjoint norms have two sources of conflict: the level of commitment and the structural conflict (beneficiary vs. target actor). Content-related conflicts over conjoint norms also have two sources of conflict: the level of commitment and the content. Finally, content-related conflicts over disjoint norms have three sources of conflict: the level of commitment, the content and the structure (beneficiary vs. target actor).

4 Experimental evidence on the effectiveness of peer punishment for different types of norm-related conflicts

In the following, we systematize experimental research on the effectiveness of punishment. This is done by discussing exemplary findings for each type of normative conflict.

4.1 Commitment-related conflicts over conjoint norms

A classic study on the effectiveness of punishment in public-good provisions is Fehr and Gächter (2002). In this study, groups of four could invest in a linear public good with a marginal *per capita* return of 0.4. This creates a situation where everybody's egoistic incentive is to contribute nothing (since every monetary unit yields individual returns of 0.4). However, if everybody contributes, everybody receives higher earnings ($4 \cdot 0.4 = 1.6$ received units from each contributed unit). This creates a conjoint cooperation norm, since group members are beneficiaries and target actors for the contribution to the public good.

In one condition, group members could punish others after having seen their contribution level. In this way, different levels of commitment to the cooperation norm could be coordinated. In most cases, high contributors punished low contributors. This increased the commitment to almost full contributions. In a condition without punishment, however, cooperation decreased substantially (Figure 3). This finding has been replicated several times and has become a textbook result in behavioral game theory (e.g., Camerer 2003). In light of our theory, the study demonstrates the high effectiveness of punishment for commitment-related disagreements about how much to contribute to a conjoint cooperation norm.

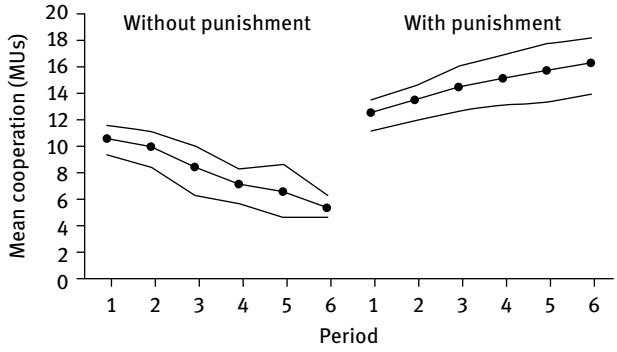


Fig. 3: Peer sanctioning enables cooperation norms in public-goods problems (Source: Fehr and Gächter 2002).

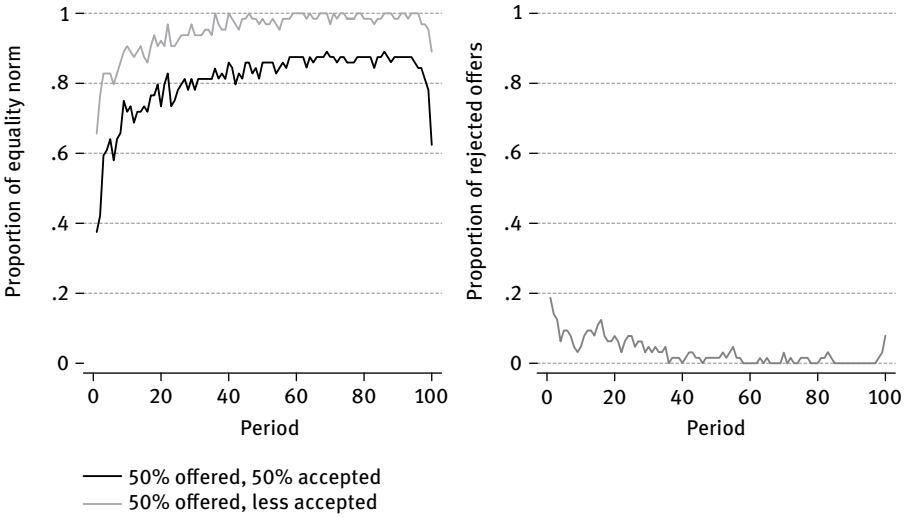
4.2 Commitment-related conflicts over disjoint norms

Disjoint norms are situations in which beneficiaries and target actors of a norm fall apart. The ultimatum game offers an abstract conceptualization of this conflict of interests. A proposer decides how much of a common pie to distribute to a responder. The responder can accept or reject. Rejection destroys all payments for both parties. The structurally stronger proposer is the target of a fairness norm to split equally (50 : 50). The structurally weaker responder benefits from this norm, because a rational and egoistic proposer would offer the smallest possible amount, which a rational and egoistic responder would accept (since this is more than nothing). Disjoint fairness norms can, however, sustain a fairness norm for two reasons (Fehr and Schmidt 1999). The proposer splits close to equal if she is inequality-averse and prefers equal outcomes compared to unequal, but higher, personal earnings. Second, the proposer could believe that the responder is sufficiently inequality-averse and prefers equal zero earnings compared to unequal positive earnings. This also generates a fairness norm. Several studies support both arguments: proposers offer substantial amounts even without a rejection possibility, and the proposers' violations of a fairness norm are often punished by the responders' rejections (Camerer 2003).

Avrahami et al. (2013) conducted an often-repeated ultimatum game experiment with changing partners. This design allows us to study the evolution of fairness norms and the effectiveness of punishment for norm enforcement. We reanalyzed their data to yield some support for our conjecture about the effectiveness of punishment. Our analysis showed that the adherence to a fairness norm of 50 : 50 quickly and strongly converges towards consensus (Figure 4 left). Violations of this norm are punished by rejections (Figure 4 right). Since the proportion of multilateral norm adherence strongly increases, the occurrence of punishment decreases over time, suggesting that the norm reproduces itself over and over again.

This experiment simulates an abstract scenario of commitment-related conflicts in disjoint norms. Mostly, the proposer and the responder agree that the proposer should offer some part of the pie to the responder. However, both can disagree about the proportion, that is, about the level of commitment to the fairness norm of a fully equal split.

We argue that in disjoint norms, the conflict of interests between beneficiary and target actor makes punishment less effective compared to conjoint norms. Some evidence for this argument can be deduced from a comparison of Figures 3 and 4. In disjoint norms (the ultimatum game), the norm takes longer to evolve and breaks down at the end. The evolution of the conjoint cooperation norm in public-good provisions is faster and there is no endgame effect (i.e., no breakdown of cooperation).



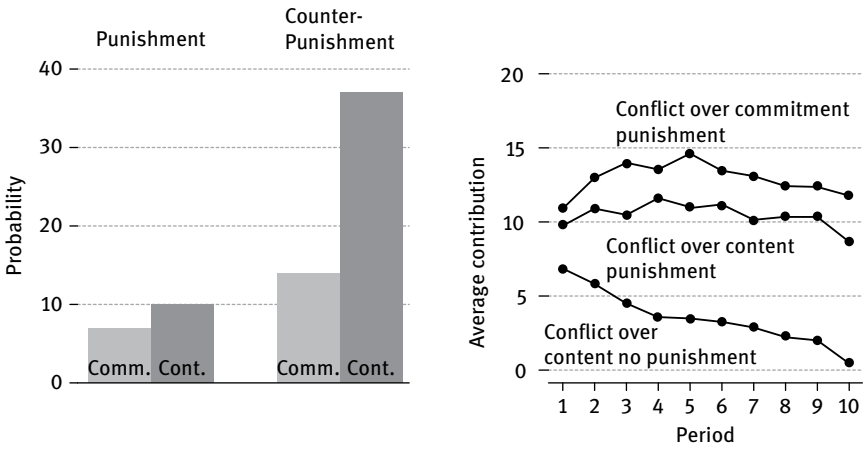
Notes: The grey line in the left panel shows the proportion of unilateral full adherence to the equality norm over time (proposer offers 50 : 50). The black line shows bilateral full adherence to the equality norm (proposer offers and responder demands 50 : 50). The right panel shows the proportion of responders' peer sanctioning of violations of the fairness norm over time (rejections by responders).

Fig. 4: Convergence of a disjoint fairness norm in a repeated ultimatum game (Source: Own compilation of reanalyzed data from Avrahami et al. 2013).

4.3 Content-related conflicts over conjoint norms

We conjecture that content-related conflicts are stronger than commitment-related conflicts. The dispute is not only about how much self-interest should be sacrificed to comply with the norm. It is a conflict about different principles, and about different conceptions of how to produce the norm. As in the argument above, we expect more conflict for disjoint than for conjoint norms.

An experimental implementation of content-related conflicts over conjoint norms is given by Nikiforakis, Noussair, and Wilkening (2012). As in the experiment by Fehr and Gächter (2002) discussed above, they designed a public-goods experiment with groups of four, where individuals could invest up to 20 monetary units in a public good in each period. The marginal *per capita* return was always such that individual contributions yielded lower returns, but collective contributions yielded higher average group earnings, creating a social dilemma. There was a baseline punishment condition like the one in Fehr and Gächter (2002). We call this condition "no normative conflict with punishment" (Figure 5). In this condition, there was a symmetric marginal *per capita* return of 0.4 for all group members. This yielded an individual return of 0.4 for each contributed unit and a 60 % group benefit from every unit contributed by oth-



Notes: The left panel shows the probability of punishment and counter-punishment in collective-goods games with only commitment-related conflicts (white) and with content-related normative conflicts (black). The right panel shows the dynamical consequences of punishment and counter-punishment in terms of average collective-good provisions. The lines refer to a symmetric game (without content conflicts; upper line), an asymmetric game (with content conflicts; middle line) and a control treatment without punishment (lower line).

Fig. 5: Normative conflict leads to feuds and less effective punishment (Source: Own compilation based on the data by Nikiforakis, Noussair, and Wilkening 2012).

ers. Extending previous experiments, counter-punishment was allowed. This means punished individuals could punish back, which could again be retaliated and so forth. Hence, feuds in terms of punishment series were allowed.

This treatment was contrasted with an asymmetric public goods game with punishment. We call this treatment "normative conflict with punishment" (Figure 5). The asymmetry was implemented in terms of different *per capita* returns. Prior to the experiment, subjects competed in a real-effort task about advantageous positions in the public goods game. Winners were selected to receive high marginal *per capita* returns (0.5), and losers were selected to receive low marginal *per capita* returns (0.3). This created a situation in which winners had higher returns from public-good contributions than losers. In this sense, winners had a stronger interest in the public good than losers.

The asymmetry in returns created normative conflicts between two possible contribution norms. First, actors could adhere to a libertarian norm and demand equal contributions from all group members (which would result in higher earnings for winners). Alternatively, actors could adhere to an equality (redistribution) norm and demand that all group members should earn equally (requiring higher contributions from winners). To put it differently, the first norm prescribed equal inputs (and implied

unequal outputs). The second norm prescribed equal outputs (and implied unequal inputs).

Both treatments were compared with a control condition in which no punishment was implemented. Otherwise, this condition was similar to the last one mentioned inasmuch as *per capita* returns were asymmetric. We call this treatment "normative conflict without punishment".

The left panel in Figure 5 shows punishment and counter-punishment probabilities for both punishment treatments. Counter-punishment is about three times as likely and about 70 % more severe in the asymmetric treatment with normative conflict over contents (black bars) compared to the symmetric treatment without normative conflict over commitments (white bars).

Counter-punishment can be regarded as an indicator of normative conflict for the following reason. If the punished party adheres to a different norm from the punisher, punishment is unjustified from the perspective of the punished party. A normatively adequate response is counter-punishment. In this sense, normative conflicts are measurable by punishment feuds.

The macro-level consequences of normative conflicts and counter-punishments are lower levels of cooperation. This is demonstrated by the right panel of Figure 5. The contributions in public-goods problems with normative conflicts (middle line) are considerably lower than in the condition without normative conflicts (upper line). This is due to more and harsher counter-punishments in the case of normative conflicts.

Both treatments can be compared to a version without the possibility of punishment (right panel, lowest line, "normative conflict, no punishment"). Without punishment, normative conflicts cannot be resolved and cooperation breaks down completely. It is noteworthy that the breakdown of cooperation is stronger than in the symmetric version without punishment, as Fehr and Gächter outlined it (2002). Kingsley (2016) replicates this finding on the adverse effects of content-related normative conflict in a similar study. More importantly, however, he showed that punishment loses its effectiveness even without the possibility of counter-punishment. Taken together, these results indicate that persistent content-related conflicts destroy cooperation more severely than commitment-related conflicts.

4.4 Content-related conflicts over disjoint norms

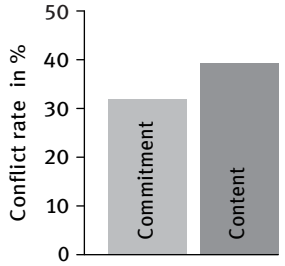
We argue that the strongest conflict with the least effective punishment is the case of content-related conflicts over disjoint norms. Here, people disagree about the normative rule and beneficiary and target actors have different interests. One example where people disagree about normative contents is when they have put different levels of effort into a collective good or have experienced different outcomes from it. A case where norms are disjoint is given if targets do not benefit from the norm. An exemplary abstract strategic setting of this kind is a heterogeneous ultimatum game.

Winter, Rauhut, and Helbing (2012) conducted such a heterogeneous ultimatum game experiment. Participants engaged in a real-effort task several days before the experiment. This yielded different monetary endowments for proposers and responders. These different endowments were based on different levels of effort. People could specify their offers to responders and their least acceptable offer from proposers for both roles (the "strategy vector method"). They were then assigned roles and partners, who typically had different endowments to contribute to the common pie.

About half of the participants acted according to an equality norm. As proposers, they offered an equal split to the responders, and as responders, they demanded an equal split. The other half of the participants, however, acted according to an equity norm. As proposers, their offers were proportional to their effort. They offered less to the responders if the responders had contributed less than themselves. Likewise, they offered more to the responders if the responders had contributed more than they had. About half of the responders followed this pattern and demanded offers that were proportional to their level of effort. This norm can be regarded as an alternative fairness rule where outcome is proportional to input.

The two different norms generate conflict if the proposer has contributed more than the responder, and if the proposer holds an equity and the responder an equality norm. In this case, the proposer offers less than half to the responder, while the responder requests half of the pie.

Winter, Rauhut, and Helbing (2012) estimated normative types (equity versus equality norm followers) and analyzed the likelihood of conflicts for pairs holding similar and different norms. Conflicts in the ultimatum game were operationalized as rejected offers. It turned out that conflicts occurred substantially more often if actors disagreed about the normative content than if they adhered to the same normative content (Figure 6). This gives evidence for our theory that content-related conflicts in disjoint norms represent the most severe case of conflict.



Notes: Conflict rates represent rejections in ultimatum games. No normative conflict represents cases where proposer and responder adhere to the same norm (both equality or both equity). Normative conflict represents cases where proposer and responder hold different norms (one equality and the other equity).

Fig. 6: Conflict rates without (content-related) normative conflicts (left) and with (content-related) normative conflicts (right) (Source: Own compilation based on the data by Winter, Rauhut, and Helbing 2012).

5 Implications

Our terminology developed here sorts our reasoning about conflicts and promotes a more multi-faceted view about the underlying mechanisms of normative behavior. Our theoretical arguments may invite people to rethink cooperation failures observed in the lab and in the field. One such example may be the explanation of seemingly "antisocial" behavior. Hermann, Thöni, and Gächter (2008) conducted a public-goods experiment with punishment (see section 4.1) in several different countries. They found that some societies tend to limit punishment to low contributors, while others also punish high contributors. The authors argued that their results might best be explained by heterogeneity in "civic duty norms" across societies: "[I]f participant pools held different social norms with regard to cooperation and free-riding, they actually might have punished differently" (Hermann, Thöni, and Gächter 2008:1365, emphasis added). In contrast, our theoretical sketch developed here suggests that "antisocial" punishment is an indicator of normative conflicts *within* societies. One subgroup of a society might try to promote a high level of commitment to the collective good and only punish under-contributors. At the same time, another group might be discouraged by other people trying to force them to do anything, even if it was in their best interest. This group perceives high-contributors as overly ambitious, vain, or even hypocritical, and fears that they raise the bar of cooperation too high. A similar norm of modesty has already been reported in the Hawthorne experiments by Roethlisberger and Dickson (2003:384 [1939]). Instead of enforcing high contributions, they punish those who contribute too much. Norm violations are thus punished by two opposed groups: over- and under-contributors.

6 Conclusion

This chapter outlines new theoretical ideas about normative conflicts and provides a new typology. Four types are distinguished based on the distinction between conjoint and disjoint norms by Coleman (1990) and our own classification of commitment-related and content-related normative conflicts (Winter, Rauhut, and Helbing 2012). We order the four types of normative conflicts according to their conflict potential and their effectiveness with which conflict can be restored by punishment.

So far, the literature discussed commitment-related conflicts as the main problem. Here, people must agree on the extent to which social norms should restrain their self-interest. Despite agreement that a specific norm should be followed, "undercutting" is regarded as legitimate by some and unacceptable by others. Thus, different degrees of normative commitment are an important source of normative conflict.

However, we conjecture that content-related conflicts are more severe than commitment-related ones. Consequently, we expect punishment in content-related con-

flicts to be less effective in restoring cooperation. Despite actors deciding to be cooperative and contributing an appropriate share to the commons, they hold different norms of what they consider to be fair.

The driving factors of commitment-related conflicts are different levels of selfishness or diverging beliefs about the cooperativeness of others. We expect people to be relatively open to persuasion to be more cooperative when others are also cooperative. We argue that people are also relatively open to argumentation that others are more cooperative than they had believed.

In contrast, we expect content-related conflicts to be less easy restorable. When actors hold distinct convictions (i.e., when there is normative conflict), different normative viewpoints tend to be strongly defended, making more conflict resolution necessary. Others must be made amenable to different points of view. Communication, clarification and approval of distinct moral principles need more time and energy and more complex kinds of conflict resolution than punishment. For example, taking turns can be one solution for the peaceful coexistence of different moral principles (Winter 2014).

An obvious next research step would be the development of an empirical research design through which all types of conflicts can be studied in a more comparable way. Our comparison over different experiments, subject pools and designs is limited to providing some insights and novel ideas. A next step would require a setup in which only the types of conflicts vary. The most direct test of our theoretical conjectures would be a laboratory design in which all types of normative conflicts were implemented and subjects were randomly allocated to different types of normative conflicts. A measure of normative conflict could be the extent of counter-punishment in all four types of normative conflicts. Ideally, such a laboratory design would go hand in hand with an analytical model, from which the hypothesized extent of conflict and effectiveness of punishment can be deduced.

Despite not having formulated a rigorous theoretical model and having not provided results from a tailor-made laboratory experiment, we believe that our typology has many new implications for the understanding of when social order emerges spontaneously and how it can be organized by mechanism design. We believe that our new perspective can guide social theory and be applied to conflict resolution in the understanding and management of social norms, cooperation, and conflicts.

Bibliography

[1] Ashworth, Tony. 1980. *Trench warfare 1914–1918: the live and let live system*. New York: Holmes & Meier.

[2] Axelrod, Robert M. 1984. *The evolution of cooperation*. New York: Basic Books.

[3] Avrahami, Judith, Werner Güth, Ralph Hertwig, Yaakov Kareev, and Hironori Otsubo. 2013. "Learning (not) to yield: An experimental study of evolving ultimatum game behavior." *The Journal of Socio-Economics* 47:47–54.
[4] Berger, Roger, and Heiko Rauhut. 2014. "Reziprozität und Reputation." Pp. 715–742 in *Handbuch Modellbildung und Simulation*, edited by N. Braun, and N. Saam. Wiesbaden: VS Verlag.
[5] Camerer, Colin. 2003. *Behavioral game theory: Experiments in strategic interaction*. Princeton: Princeton University Press.
[6] Coleman, James S. 1990. *Foundations of social theory*. Cambridge, MA: The Belknap Press of Harvard University Press.
[7] Dahrendorf, Ralf. 1958. *Homo Sociologicus. Ein Versuch zur Geschichte, Bedeutung und Kritik der Kategorie der sozialen Rolle.* Opladen: Westdeutscher Verlag.
[8] Diekmann, Andreas, and Peter Preisendörfer. 2003. "The behavioral effects of environmental attitudes in low-cost and high-cost situations." *Rationality and Society* 15(4):441–472.
[9] Diekmann, Andreas. 2010. "Not the First Digit! Using Benford's Law to Detect Fraudulent Scientific Data." *Journal of Applied Statistics* 34(3):321–329.
[10] Durkheim, Emile. [1897] 1997. *Suicide*. Glencoe, IL: Free Press.
[11] Elster, Jon. 1989. *The Cement of Society: A Study of Social Order*. Cambridge: Cambridge University Press.
[12] Fehr, Ernst, and Simon Gächter. 2000. "Fairness and Retaliation: The Economics of Reciprocity." *Journal of Economic Perspectives* 14:159–181.
[13] Fehr, Ernst, and Simon Gächter. 2002. "Altruistic Punishment in Humans." *Nature* 415(10):137–140.
[14] Fehr, Ernst, and Klaus M. Schmidt. 1999. "A Theory of Fairness, Competition and Cooperation." *Quarterly Journal of Economics* 114(3):817–868.
[15] Gambetta, Diego. 2009. "Signaling." Pp. 168–194 in *The Oxford Handbook of Analytical Sociology,* edited by P. Hedström, and P. Bearman. Oxford: Oxford University Press.
[16] Hechter, Michael, and Karl-Dieter Opp. 2001. "Introduction." Pp. xi–xx in *Social norms,* edited by M. Hechter, and K.-D. Opp. New York: Russell Sage Foundation.
[17] Herrmann, Benedikt, Christian Thöni, and Simon Gächter. 2008. "Antisocial punishment across societies." *Science* 319(5868):1362–1367.
[18] Hyman, Richard, and Ian Brough. 1975. *Social Values and Industrial Relations: Study of Fairness and Inequality (Warwick Studies in Industrial Relations)*. Oxford: Blackwell Publishers.
[19] Kingsley, David C. 2016. "Endowment heterogeneity and peer punishment in a public good experiment: Cooperation and normative conflict." *Journal of Behavioral and Experimental Economics* (in press).
[20] Nikiforakis, Nikos, Charles N. Noussair, and Tom Wilkening. 2012. "Normative conflict and feuds: The limits of self-enforcement". *Journal of Public Economics* 96(9):797–807.
[21] Nowak, Martin A., and Karl Sigmund. 1998. "Evolution of Indirect Reciprocity by Image Scoring." *Nature* 393(6685):573–577.
[22] Olson, Mancur. 1965. *The Logic of Collective Action*. Cambridge, MA: Harvard University Press.
[23] Parsons, Talcot. 1937. *The Structure of Social Action*. Glencoe, IL: Free Press.
[24] Röthlisberger, Fritz J., and William J. Dickson. 2003. *The early sociology of management and organizations*. Vol. 5, *Management and the Worker*. London and New York: Routledge.
[25] Przepiorka, Wojtek, and Diekmann, Andreas. 2013. "Individual heterogeneity and costly punishment: a volunteer's dilemma." *Proceedings of the Royal Society of London B: Biological Sciences* 280(1759):20130247.
[26] Sigmund, Karl. 2010. *The Calculus of Selfishness*. Princeton, NJ: Princeton University Press.
[27] Stouffer, Samuel A. 1949. *The American Soldier*. Princeton, NJ: Princeton University Press.
[28] Ullmann-Margalit, Edna. 1977. *The Emergence of Norms.* Oxford: Clarendon Press.

[29] Voss, Thomas. 2001. "Game-Theoretical Perspectives on the Emergence of Social Norms." Pp. 104–136 in *Social norms*, edited by M. Hechter, and K.-D. Opp. New York, NY: Russell Sage Foundation.
[30] Wedekind, Claus, and Manfred Milinski. 2000. "Cooperation Through Image Scoring in Humans." *Science* 288(5467):850–852.
[31] Winter, Fabian, Heiko Rauhut, and Dirk Helbing. 2012. "How norms can generate conflict: An experiment on the failure of cooperative micro-motives on the macro-level." *Social Forces* 90(3):919–948.
[32] Winter, Fabian. 2014. "Fairness Norms Can Explain the Emergence of Specific Cooperation Norms in the Battle of the Prisoner's Dilemma." *The Journal of Mathematical Sociology* 38(4):302–320.